

German Patent No. DE 10,062,566 A1

Job No.: 6592-92265

Ref.: DE 10062566

Translated from German by the Ralph McElroy Translation Company
910 West Avenue, Austin, Texas 78701 USA

FEDERAL REPUBLIC OF GERMANY
GERMAN PATENT OFFICE
PATENT NO. DE 100 62 566 A1
(Offenlegungsschrift)

Int. Cl.⁷: C 12 Q 1/68
Filing No.: 100 62 566.5
Filing Date: December 15, 2000
Date Laid-Open to Public Inspection: June 20, 2002

METHOD AND DEVICE FOR DETERMINING PRIMER SEQUENCES

Inventor(s): Oliver Hummel
13189 Berlin, Germany
Applicant(s): GenProfile AG
13125 Berlin, Germany
Agent: v. Bezold & Sozien
80799 Munich, Germany

The following statements are taken [unedited] from the documents submitted by the applicant.

A method for determining primer sequences and primer pairs for the production of a plurality of neighboring or overlapping PCR fragments for a given DNA sequence comprises the following steps: Listing a plurality of segments of the sequence which, based on their length, can be selected as primers; determination of the values of characteristic parameters for each segment, where a minimum of one parameter in the transition regions between parameter value ranges, for which the given parameter is considered optimum or unsuitable for producing PCR fragments, is continuously evaluated by means of a continuous and strictly monotonic transition function, with the given parameter value being assigned a value in correspondence with the quantity of the associated transition function; exclusion of segments which, with respect to a minimum of one parameter, are considered unsuitable for the production of PCR fragments; overall evaluation of the remaining segments by means of a weighted sum formation of the quantities assigned to the parameter values during the continuous evaluation; selection of the segments -- which, based on

the overall evaluation, correspond to predetermined selection criteria -- as primers; and storage, display and/or readout of the selected segments as primer sequences and grouping into primer pairs.

//insert Figure//

Description

[0001]

The subject matter of the present invention relates to methods for determining primer sequences, in particular methods for selecting primers for a given DNA sequence on the basis of a continuous evaluation of characteristic primer parameters, devices for carrying out the methods, and uses thereof.

[0002]

Sequencing, i.e., determining the order of nucleotides in nucleic acid sequences, is the basic method for deciphering the genetic code. In addition to identifying unknown sequences, sequencing also serves, among other things, to reveal individual differences in known sequences. When sequencing long DNA sequences, it is necessary, for technical reasons, to divide the sequences into manageable fragments, the so-called PCR fragments. These fragments are amplified by means of the polymerase chain reaction with highly specific oligonucleotides, the so-called primers. The fragments are bounded and determined by one primer pair on each side.

[0003]

The selection of the suitable primer pair -- in particular the distance between the two primers in the sequence to be amplified, from which distance the length of the fragments results -- is important for the properties of a PCR fragment. In addition, multiple binding sites of one or both primers in the sequence must be taken into consideration so as to be able to exclude potential by-products. To eliminate potential sequencing errors at the ends of the fragments, neighboring fragments with a minimum overlap of, e.g., 100 bases can be generated. Primers are highly specific reagents. Their quality depends on a plurality of parameters. Some of these

parameters are interdependent on one another: e.g., length, melting temperature, GC content, 3' pentamer stability, frequency of the 3' heptamer, and ability to form dimers or hairpins.

[0004]

There are several known computer programs which can be used to select high-quality primers for individual PCR fragments: e.g., "Primer3" (S. Rozen et al. in "Methods Mol. Biol." 132, 2000, pp. 365-386) or prime+ (GCG - Genetics Computer Group, 1998), Program Manual for the GCG Package, Version 10, 1998, 575 Science Drive, Madison, Wisconsin 53711, USA). These programs, however, are not designed to efficiently generate overlapping PCR fragments for sequencing very long genome segments.

[0005]

In the method implemented with "Primer3," a plurality of segments of a predetermined sequence are tested with respect to a group of predetermined characteristic primer parameters. For each parameter, a range of values or/and an optimum value is/are defined, within which range or optimum value segments of the DNA sequence are suitable as primers. Values outside the range/optimum value are penalized by means of a reduction of points. The selection of a relatively large number of PCR primers is inefficient because of the discontinuous evaluation which acts as an excessively restrictive filter.

[0006]

Thus, the problem to be solved by the present invention is to make available a new method for determining primer sequences by means of which the above-mentioned disadvantages of the conventional methods can be overcome and which makes possible an automated selection of PCR primers. Another objective of the present invention is to make devices available for implementing the methods as well as uses of the methods.

[0007]

These problems are solved by means of the methods, computer program products or devices with the features according to Claims 1, 11, and 12. Useful embodiments of this invention follow from the dependent claims.

[0008]

The basic idea of this invention is to cover the DNA sequence on the basis of a continuous evaluation and a weighted overall evaluation of the primer parameters with optimum primer pairs. In addition to the conventional parameters (see above), it is possible to include any

other parameters as well. Because of the continuous evaluation and weighting of all parameters under consideration, the method according to the present invention optimizes the primer quality for automated sequencing.

[0009]

The present invention has the advantage that the suitability of a segment to serve as a primer is evaluated more objectively and more realistically than is possible with the conventional techniques and that the analysis of long DNA sequences in the form of overlapping sequence segments as units for analysis is made possible in an efficient manner. This makes possible a high analytical throughput at a low downtime rate which results in savings in material, labor, and time.

[0010]

Another subject matter of the present invention relates to a computer program for implementing the method described.

[0011]

Yet another subject matter of the present invention relates to a device for determining primer sequences with a means for identifying a plurality of segments of a given sequence and for evaluating the characteristic primer parameters, a first selection means which is designed to dispose of segments of the sequence that are unsuitable as primer sequences, a second selection means which is designed to compare the sum of the individual weighted evaluations of the remaining segments with a predetermined selection criterion, e.g., an absolute threshold value and/or a maximum value formation, and a third selection means which is designed to determine the optimum primer pairs. The device according to the present invention also comprises means for the storage, the readout and/or the display of the determined primer sequences and primer pairs.

[0012]

The method according to the present invention can be usefully employed in many applications in biology, biotechnology, human and veterinary medicine, agricultural sciences, ecology, and all other fields in which the sequencing of nucleic acids comes into play.

[0013]

Additional advantages and details of the present invention follow from the explanation of the process sequence, the description of an example, and the enclosed drawing. Figure 1 shows an example of a continuous transition function used according to the present invention to

evaluate a primer parameter (x-axis: parameter values, y axis: evaluation of the parameter values).

Detailed description of the method according to the present invention

[0014]

The objective of the present invention is to automatically select PCR primers with a continuous evaluation of the parameters on the basis of a randomly long, known sequence (the sequence which is obtained, e.g., from the EMBL (European Molecular Biology Laboratories) database and by means of predetermined parameters. For this purpose, first the entire length or a suitable fragment of the sequence is characterized for all possible fragments along this length with respect to the primer parameters of interest. The characteristic parameters comprise specific properties of the primer (e.g., primer length, GC content, melting temperature of the primer, etc.) as well as parameters which characterize predetermined criteria (e.g., 3' pentamer stability, frequency of the 3' heptamer, or the occurrence of repetitive sequences ("repeats").

[0015]

The fragment lengths and the minimum overlapping range of the individual fragments are pre-specified as a function of other factors, e.g., the sequencing method, the reagents used, and the length of the DNA segment to be sequenced. In addition, regions of the DNA to be sequenced can be "masked" if no PCR primers are to be selected from them. These may be sequence regions which occur frequently in the genome (so-called "repeats") or known polymorphisms and other sites that are unsuitable for the selection of the primers (e.g., within an exon). On the basis of all of the given specifications, the optimum primers are determined.

1. Characterization of the complete sequence

[0016]

In a first step, the complete sequence is tested with respect to the included parameters for all potential primer positions and lengths. Among other things, repeat regions, known polymorphisms, and other sequence regions that are unsuitable for the primer selection are masked. In a second step, the values of the included parameters (e.g., GC content, melting temperature, 3' pentamer stability, etc.) are determined for all predetermined primer lengths. Subsequently, a value between 0 and 1 is assigned to each parameter on the basis of a parameter-specific continuous or discontinuous transition function.

[0017]

According to the present invention, the transition functions can assume continuous values between 0 and 1. "1" and "0" can also be described, respectively, as "optimum suitability" and "no suitability" as primers with respect to the concrete property. One example of this is illustrated in Figure 1 for the primer GC content (guanosine plus cytosine content). To values above and below predetermined upper and lower threshold values A and D, respectively, 0 is assigned. Primers with a parameter to which 0 was assigned may not be selected. Within the interval that is bounded by the threshold values A and D, there are two other threshold values (B, C). To the enclosed interval from B to C, 1 is assigned, i.e., the parameter is optimally suitable within the range of values from B to C. The optimum for each parameter may cover a relatively wide range of values. The optimum is flanked by regions (the region between A and B and that between C and D) in which the quality of the potential primer for the parameter under consideration decreases. In these regions, a continuous evaluation is performed by means of a minimum of one transition function. This decrease can be characterized by the most varied functions, which can differ even from one parameter to the next.

[0018]

As illustrated in Figure 1, to mathematically describe the evaluation of a parameter in these two regions, e.g., linear functions and trigonometric functions are suitable. As a result, this makes possible a continuous evaluation of all or of selected parameters which characterize a primer. In contrast to the already available methods of primer designs, the evaluation of the parameters does not take place discontinuously.

2. Overall evaluation

[0019]

Based on the individual evaluations of all parameters of a primer thus obtained, an overall evaluation is made. Once an included parameter of a primer has been assigned the value 0, this primer is excluded from the selection. For all remaining potential primers, the overall evaluation is made by means of a summation of the evaluations of all included parameters, with the possibility of also freely selecting a weighting of each individual parameter. To determine the optimally suitable primer pairs, a weighted continuous evaluation of the fragment length (which results from the positions of the forward and backward primers in the sequence) and of the two melting temperatures of the primers is carried out in addition to the overall evaluation of the two individual primers. For this purpose, a continuous evaluation of the difference in the melting temperature between the forward and the backward primer is made.

[0020]

The primer pair with the highest sum of weighted evaluations is selected. In addition to the 5' end of the sequence, critical, e.g., masked, regions can be used as the starting point for the selection of primer pairs as well.

Example

1. Selection of the characteristic parameters and the transition functions thereof

[0021]

The objective is to determine primers for the sequence with the Accession Number "D26607" from the EMBL database ("<http://www.ebi.ac.uk/cgi-bin/emblfetch?D26607>"). The sequence has a length of 23,142 bp (base pairs). Primers for the region between base pair 1 and 4000 are to be sampled. The PCR fragments are to have a length of 1050 bp. The overlap of the fragments should be at least 100 bp. The primers are to have a length of 18 to 27 nucleotides. If it is not possible to generate fragments, the length of which differs at most by 120 bp from the predetermined fragment length, the predetermined fragment length is to be increased or decreased by 350 bp. The following parameters are included as the characteristic parameters of the primers:

- Length
- GC content
- Melting temperature
- Number of 3' heptamer in the complete sequence
- Free energy of the 3' pentamer
- Masking
- Belonging to an exon
- Occurrence of the 3' heptamer within a region of 10 kbp.

[0022]

The last three parameters can assume only the values 0 or 1. All other parameters are continuously evaluated. Primers for which the value 0 is assigned to a minimum of one parameter are excluded from the subsequent calculations.

[0023]

or the evaluation of the parameter values between points A and B as well as between C and D, two functions are selected (see Figure 1):

- for region [A, B[:

Function 1:

//insert A//

Function 2:

//insert B//

- for region]C, D]:

Function 3:

//insert C//

Claims

1. A method for determining primer sequences and primer pairs for the production of a plurality of neighboring or overlapping PCR fragments for a given DNA sequence, comprising the following steps:

- Preparation of a list with a plurality of segments of the given sequence which, on the basis of their length, can be selected for use as primers,
- Determination of the values of characteristic parameters for each segment, where a minimum of one parameter is continuously evaluated by identifying parameter value ranges which are characteristic for the parameter and for which the given parameter is considered either optimum or unsuitable for the production of PCR fragments and where these parameter value ranges are separated by transition regions which are represented by transition functions that are continuous and which follow a strictly monotonic course, with an evaluation corresponding to the quantity of the associated transition function being assigned to the respective parameter value in the transition regions,
- Exclusion of the segments which, with respect to a minimum of one parameter, are unsuitable for use in the production of PCR fragments,

- Overall evaluation of the remaining segments by means of a weighted sum formation of the quantities assigned to the parameter values during the continuous evaluation,
- Selection of the segments, which, based on the overall evaluation, correspond to predetermined selection criteria, as primers and
- Storage, display and/or readout of the selected segments as primer sequences and grouping into primer pairs.

2. The method as claimed in Claim 1, in which the position of the segment in the sequence, the length of the segment, the GC content, the melting temperature, the 3' oligomer stability, the 3' oligomer frequency and/or the masking of unsuitable sequence regions are determined as the parameters for the segment.

3. The method as claimed in Claim 2, in which segments with lengths of less than 12 and more than 35 nucleotides are evaluated as being unsuitable and segments with lengths of 20 to 24 nucleotides are evaluated as being optimum.

4. The method as claimed in Claim 2 or 3, in which segments with a GC content of lower than 15% and higher than 85% are evaluated as being unsuitable and those with 45% to 55% are evaluated as being optimum.

5. The method as claimed in one of Claims 2 through 4, in which segments with a melting temperature lower than 30°C and higher than 90°C are evaluated as being unsuitable and those in a temperature range between 50°C and 55°C are evaluated as being optimum.

6. The method as claimed in any one of the preceding claims, in which the transition functions are formed by trigonometric functions, linear functions, or overlaps thereof.

7. The method as claimed in any one of the preceding claims, in which the selection of segments as primers with respect to the fragment length takes place so that, given a predetermined optimum length of segments for a first selected segment on a strand of the sequence, the segments on the other strand are evaluated as being unsuitable if they lead to a fragment length which is lower than a predetermined minimum length or higher than a predetermined maximum length.

8. The method as claimed in any one of the preceding claims, in which the selection of segments as primers takes place so that for a first selected segment on the one strand of the sequence, the segments on the other strand are evaluated as being unsuitable if both segments show a difference in the melting temperature that is higher than 10°C, with the segments being evaluated as being optimum if the difference in the melting temperature is lower than 0.5°C.

9. The method as claimed in Claim 8, in which, in a sequence region for which a segment on the one strand of the sequence is to be selected as a primer, the 15 primers with the highest overall evaluations of the individual primers are pre-selected and for which, in correspondence with the predetermined optimum fragment length, the optimum associated segments on the

oppositely lying strand are selected, with the pair with the highest overall evaluation -- that results from the evaluation of the individual segments, the difference of the melting temperature between the involved primers, and the fragment length -- being selected from the pairs of segments thus formed.

10. The method as claimed in Claim 9, in which, within a predetermined region of the sequence, the position of the primer relative to the neighboring fragment is identified as an additional parameter so as to form a predetermined minimum overlap.

11. A computer program product which is designed to determine primer sequences and primer pairs using a method according to one of the preceding claims.

12. A device for determining the primer sequences and primer pairs, using a method according to one of the preceding claims and comprising:

- a means for determining a plurality of segments of a given DNA sequence and of associated characteristic parameters,
- an evaluation means for evaluating the determined parameters,
- a first selection means which is designed to dispose of segments of the sequences that are unsuitable as primers,
- a second selection means which is designed to evaluate the remaining segments, and
- means for the storage, readout and/or display of the primer sequences and primer pairs determined.

Includes 1 page of drawings

⑯ Aktenzeichen: 100 62 566.5
⑯ Anmeldetag: 15. 12. 2000
⑯ Offenlegungstag: 20. 6. 2002

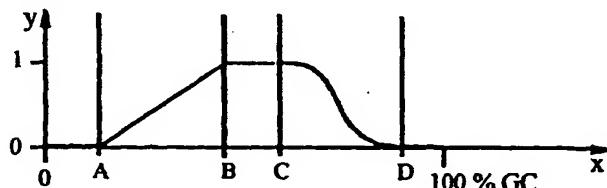
⑯ Anmelder:
GenProfile AG, 13125 Berlin, DE
⑯ Vertreter:
v. Bezold & Sozien, 80799 München

⑯ Erfinder:
Hummel, Oliver, 13189 Berlin, DE

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

⑯ Verfahren und Vorrichtung zur Ermittlung von Primer-Sequenzen

⑯ Ein Verfahren zur Ermittlung von Primer-Sequenzen und Primer-Paaren zur Erzeugung einer Vielzahl benachbarter oder überlappender PCR-Fragmente für eine gegebene DNA-Sequenz umfasst die Schritte: Auflistung einer Vielzahl von Teilstücken der Sequenz, die aufgrund ihrer Länge als Primer ausgewählt werden können; Ermittlung der Werte von charakteristischen Parametern für jeden Teilstück, wobei mindestens ein Parameter in Übergangsbereichen zwischen Parameterwertebereichen, für die der jeweilige Parameter als optimal oder als ungeeignet zur Erzeugung von PCR-Fragmenten gilt, mit einer stetigen und streng monotonen Übergangsfunktion kontinuierlich bewertet wird, wobei dem jeweiligen Parameterwert eine Bewertung entsprechend der Größe der zugehörigen Übergangsfunktion zugeordnet wird, Ausschluss von Teilstücken, die in Bezug auf mindestens einen Parameter als ungeeignet zur Erzeugung von PCR-Fragmenten bewertet werden; Gesamtbewertung der übrigen Teilstücken durch gewichtete Summenbildung der den Parameterwerten bei der kontinuierlichen Bewertung zugeordneten Größen; Auswahl der Teilstücken als Primer, die nach der Gesamtbewertung vorbestimmten Auswahlkriterien entsprechen; und Speicherung, Anzeige und/oder Ausgabe der ausgewählten Teilstücken als Primer-Sequenzen und Anordnung zu Primer-Paaren.



Beschreibung

[0001] Die Erfindung betrifft Verfahren zur Ermittlung von Primer-Sequenzen, insbesondere Verfahren zur Primer-Auswahl für eine vorgegebene DNA-Sequenz auf der Grundlage einer kontinuierlichen Bewertung charakteristischer 5 Primer-Parameter, Vorrichtungen zur Durchführung der Verfahren und deren Anwendungen.

[0002] Die Sequenzierung, d. h. die Ermittlung der Nukleotidabfolge in Nukleinsäure-Sequenzen, ist das fundamentale Verfahren zur Entschlüsselung des genetischen Codes. Neben der Aufklärung unbekannter Sequenzen dient die Sequenzierung u. a. dazu, individuelle Unterschiede in bekannten Sequenzen zu ermitteln. Zur Sequenzierung von langen DNA-Sequenzen ist es aus technischen Gründen erforderlich, die Sequenzen in handhabbare Unterabschnitte, die PCR-Fragmente, zu unterteilen. Diese Fragmente werden mittels Polymerase-Kettenreaktion mit hochspezifischen Oligonukleotiden, den sogenannten Primern, amplifiziert. Die Fragmente werden durch jeweils ein Primer-Paar begrenzt und bestimmt.

[0003] Für die Eigenschaften eines PCR-Fragments spielt die Auswahl des entsprechenden Primer-Paares eine wesentliche Rolle, insbesondere der Abstand der beiden Primer auf der zu amplifizierenden Sequenz, aus dem sich die Fragmentlänge ergibt. Außerdem müssen multiple Bindungsorte eines oder beider Primer auf der Sequenz berücksichtigt werden, um mögliche Nebenprodukte auszuschließen. Um mögliche Sequenzierungsschäler an den Enden der Fragmente zu eliminieren, können benachbarte Fragmente mit einer Mindestüberlappung von z. B. 100 Basen generiert werden. Primer stellen hochspezifische Reagenzien dar. Ihre Qualität hängt von einer Vielzahl von Parametern ab. Diese Parameter sind teilweise voneinander abhängig: z. B. Länge, Schmelztemperatur, GC-Gehalt, 3'-Pentamerstabilität, Häufigkeit des 3'-Heptamers und Fähigkeit zur Dimer- oder Hairpinbildung.

[0004] Es sind zahlreiche Computerprogramme bekannt, mit denen Primer in guter Qualität für einzelne PCR-Fragmente ausgewählt werden können: z. B. "Primer3" (S. Rozen et al. in "Methods Mol. Biol.", 2000, 132, 365-86) oder prime+ (GCG - Genetics Computer Group, (1998), Program Manual for the GCG Package, Version 10, 1998, 575 Science Drive, Madison, Wisconsin, USA 53711). Diese Programme sind allerdings nicht für die effiziente Generierung 20 überlappender PCR-Fragmente zur Sequenzierung sehr langer Genomabschnitte ausgelegt.

[0005] Bei dem von "Primer3" umgesetzten Verfahren wird eine Vielzahl von Teilstücken einer vorgegebenen Sequenz in Bezug auf eine Gruppe vorgegebener charakteristischer Primer-Parameter untersucht. Für jeden Parameter wird ein Wertebereich oder/und ein Optimalwert definiert, innerhalb dessen Teilstücke der DNA-Sequenz als Primer geeignet sind. Werte außerhalb des Bereichs/Optimalwerts werden mit einem Punktabzug bestraft. Die Auswahl einer größeren Anzahl von PCR-Primern ist wegen der diskontinuierlichen Bewertung, die einen zu strengen Filter darstellt, uneffektiv.

[0006] Die Aufgabe der Erfindung ist es, ein neues Verfahren zur Ermittlung von Primer-Sequenzen anzugeben, mit dem die oben erläuterten Nachteile der herkömmlichen Verfahren überwunden werden und das eine automatisierte Auswahl von PCR-Primern ermöglicht. Die Aufgabe der Erfindung ist es auch, Vorrichtungen zur Umsetzung der Verfahren 35 und Anwendungen der Verfahren bereitzustellen.

[0007] Diese Aufgaben werden mit Verfahren, Computerprogrammprodukten oder Vorrichtungen mit den Merkmalen gemäß den Patentansprüchen 1, 11 und 12 gelöst. Vorteilhafte Ausführungsformen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

[0008] Die Grundidee der Erfindung ist es, die DNA-Sequenz auf der Grundlage einer kontinuierlichen Bewertung sowie einer gewichteten Gesamtbewertung der Primer-Parameter mit besten Primer-Paaren abzudecken. Als Parameter können neben den gebräuchlichen (s. o.) auch jeder weitere Parameter einbezogen werden. Das erfindungsgemäße Verfahren optimiert die Primerqualität für die automatisierte Sequenzierung durch kontinuierliche Bewertung und Wichtung aller betrachteten Parameter.

[0009] Die Erfindung besitzt den Vorteil, dass die Eignung eines Teilstückes als Primer objektiver bzw. realistischer 45 bewertet wird als bei den herkömmlichen Techniken und dass die Analyse von langen DNA-Sequenzen in Form überlappender Sequenzabschnitte als Analyseinheiten in rationaler Weise ermöglicht wird. Damit wird ein hoher Analyse-durchsatz bei einer geringen Ausfallquote erzielt, woraus sich eine Einsparung an Material, Arbeit und Zeit ergibt.

[0010] Gegenstand der Erfindung ist ferner ein Computerprogramm zur Umsetzung des genannten Verfahrens.

[0011] Gegenstand der Erfindung ist auch eine Vorrichtung zur Ermittlung von Primer-Sequenzen mit einer Einrich- 50 tung zur Erfassung einer Vielzahl von Teilstücken einer gegebenen Sequenz und zur Bewertung der charakteristischen Primer-Parameter, einer ersten Auswahlvorrichtung, die dazu ausgelegt ist, als Primer-Sequenzen ungeeignete Teilstücke der Sequenz zu verwerfen, einer zweiten Auswahlvorrichtung, die dazu ausgelegt ist, die Summe der gewichteten Einzelbewertungen der übrigen Teilstücke mit einem vorbestimmten Auswahlkriterium, z. B. einem Absoluteschwellwert und/oder einer Maximalwertbildung, zu vergleichen, und einer dritten Auswahlvorrichtung, die dazu ausgelegt ist, die besten Primer-Paare zu ermitteln. Die erfindungsgemäße Vorrichtung umfasst auch Einrichtungen zur Speicherung, Ausgabe und/oder Anzeige der ermittelten Primer-Sequenzen und Primer-Paare.

[0012] Das erfindungsgemäße Verfahren besitzt vorteilhafterweise zahlreiche Anwendungen in der Biologie, Biotechnologie, Human- und Veterinärmedizin, Agrarwirtschaft, Ökologie sowie alle anderen Gebieten, in denen die Sequenzierung von Nukleinsäuren eine Rolle spielt.

[0013] Weitere Vorteile und Einzelheiten der Erfindung werden aus der folgenden Beschreibung des Verfahrensablaufs, der Erläuterung eines Beispiels und der beigefügten Zeichnung ersichtlich. Fig. 1 zeigt ein Beispiel für eine erfindungsgemäß verwendete kontinuierliche Übergangsfunktion zur Bewertung eines Primer-Parameters (x-Achse: Parameterwerte, y-Achse: Bewertung der Parameterwerte).

[0014] Erfindungsgemäß ist insbesondere vorgesehen, dass auf Basis einer beliebig langen, bekannten Sequenz (der Sequenz, die z. B. aus der EMBL (European Molecular Biology Laboratories)-Datenbank stammt) und anhand festge-

DE 100 62 566 A 1

leger Parameter PCR-Primer automatisch mit einer kontinuierlichen Bewertung der Parameter ausgewählt werden. Dazu wird die Sequenz zunächst über ihre gesamte Länge oder ein entsprechendes Teilstück für alle möglichen Teilstücke auf dieser Länge in Bezug auf die interessierenden Primer-Parameter charakterisiert. Die charakteristischen Parameter umfassen bestimmte Eigenschaften des Primers (z. B. Primerlänge, GC-Gehalt, Primer-Schmelztemperatur usw.) und Parameter, die festgelegte Kriterien charakterisieren (z. B. 3'-Pentamerstabilität, Häufigkeit des 3'-Heptamers oder das Vorkommen von repetitiven Sequenzen ("repeats")).

[0015] In Abhängigkeit von anderen Faktoren, z. B. der Sequenziermethode, den verwendeten Reagenzien und der Länge des zu sequenzierenden DNA-Abschnitts werden die Fragmentlängen und der Mindestüberlappungsbereich der einzelnen Fragmente vorgegeben. Außerdem können Bereiche der zu sequenzierenden DNA "maskiert" werden, wenn aus diesen keine PCR-Primer ausgewählt werden sollen. Dies können Sequenzbereiche sein, die im Genom häufig vorkommen (sogenannte "repeats") oder bekannte Polymorphismen sowie andere für die Primer-Auswahl ungeeignete Orte (z. B. innerhalb eines Exons). Auf der Grundlage aller festgelegten Vorgaben werden die optimalen Primer ermittelt.

1. Charakterisierung der gesamten Sequenz

[0016] In einem ersten Schritt wird die komplette Sequenz hinsichtlich der einbezogenen Parameter für alle potentiellen Primerpositionen und -längen untersucht. Es werden u. a. repeat-Regionen, bekannte Polymorphismen sowie weitere für die Primer-Auswahl ungeeignete Sequenzbereiche maskiert. Für alle vorgegebenen Primerlängen werden in einem zweiten Schritt die Werte der einbezogenen Parameter bestimmt (z. B. GC-Gehalt, Schmelztemperatur, 3'-Pentamerstabilität u. a.). Anschließend wird jedem Parameter auf der Grundlage einer parameterspezifischen kontinuierlichen oder diskontinuierlichen Übergangsfunktion eine Bewertung zwischen 0 und 1 zugeordnet.

[0017] Die Übergangsfunktionen können erfahrungsgemäß einen zwischen 0 und 1 kontinuierlichen Werteverlauf besitzen. "1" und "0" können auch als "beste Eignung" bzw. "keine Eignung" als Primer hinsichtlich der konkreten Eigenschaft beschrieben werden. Ein Beispiel hierfür ist in Fig. 1 für den Primer-GC-Gehalt (Guanosin-Cytosin-Gehalt) illustriert. Für Werte oberhalb und unterhalb vorbestimmter oberer und unterer Grenzwerte A bzw. D erfolgt die Bewertung mit 0. Primer mit einem mit 0 bewerteten Parameter dürfen nicht ausgewählt werden. Innerhalb des durch die Grenzwerte A und D aufgespannten Intervalls sind zwei weitere Grenzwerte (B, C) vorgesehen. Im abgeschlossenen Intervall von B bis C erfolgt die Bewertung mit 1, d. h. der Parameter ist im Wertebereich von B bis C optimal geeignet. Das Optimum für einen jeden Parameter darf einen relativ großen Wertebereich umfassen. Das Optimum wird von Bereichen flankiert (Bereich zwischen A und B sowie zwischen C und D), in denen die Qualität des möglichen Primers für den betrachteten Parameter abfällt. In diesen Bereichen wird eine kontinuierliche Bewertung durch mindestens eine Übergangsfunktion gebildet. Dieser Abfall kann durch unterschiedlichste Funktionen, auch verschieden für unterschiedliche Parameter, charakterisiert werden.

[0018] Zur mathematischen Beschreibung der Bewertung eines Parameters in diesen beiden Bereichen eignen sich u. a. lineare Funktionen und Winkelfunktionen, wie in Fig. 1 dargestellt. Dadurch ist eine kontinuierliche Bewertung aller bzw. ausgewählter Parameter, die einen Primer charakterisieren, möglich. Im Gegensatz zu den bereits verfügbaren Verfahren zum Primer-Design erfolgt die Bewertung der Parameter nicht sprunghaft.

2. Gesamtbewertung

[0019] Anhand der so gewonnenen Einzelbewertungen aller Parameter eines Primers wird eine Gesamtbewertung durchgeführt. Ist ein einbezogener Parameter eines Primers mit 0 bewertet worden, so wird dieser Primer von der Auswahl ausgeschlossen. Für alle übrigen möglichen Primer erfolgt die Gesamtbewertung durch Summation der Bewertungen aller einbezogenen Parameter, wobei auch eine Wichtung jedes einzelnen Parameters frei gewählt werden kann. Zur Bestimmung der am besten geeigneten Primer-Paare wird neben der Gesamtbewertung der beiden einzelnen Primer zusätzlich eine gewichtete kontinuierliche Bewertung der Fragmentlänge (die aus den Positionen von Vorwärts- und Rückwärtsprimer auf der Sequenz resultiert) und der beiden Primer-Schmelztemperaturen vorgenommen. Dazu erfolgt eine kontinuierliche Bewertung der Schmelztemperatur-Differenz zwischen Vorwärts- und Rückwärtsprimer.

[0020] Das Primer-Paar mit der höchsten Summe der gewichteten Bewertungen wird ausgewählt. Ausgangspunkte zur Auswahl von Primer-Paaren können neben dem 5'-Ende der Sequenz auch kritische, z. B. maskierte Bereiche sein.

Beispiel

1. Auswahl der charakteristischen Parameter und deren Übergangsfunktionen

[0021] Es sind Primer für die Sequenz mit der Accession Number "D26607" aus der EMBL-Datenbank ("http://www.ebi.ac.uk/cgi-bin/emblfetch?D26607") zu ermitteln. Die Sequenz hat eine Länge von 23142 bp (Basenpaaren). Es sollen Primer für den Bereich zwischen dem Basenpaar 1 und 4000 exemplarisch ausgewählt werden. Die PCR-Fragmente sollen eine Länge von 1050 bp haben. Die Überlappung der Fragmente soll mindestens 100 bp betragen. Die Primer sollen eine Länge von 18 bis 27 Nukleotiden haben. Wenn keine Fragmente generiert werden können, deren Länge höchstens um 120 bp von der vorgegebenen Fragmentlänge abweicht, soll die vorgegebene Fragmentlänge um 350 bp erhöht bzw. verringert werden. Als charakteristische Parameter der Primer werden die folgenden einbezogen:

- Länge,
- GC-Gehalt,
- Schmelztemperatur,
- Anzahl des 3'-Heptamers in der Gesamtsequenz,
- Freie Energie des 3'-Pentamers,

DE 100 62 566 A 1

- Maskierung,
- Zugehörigkeit zu einem Exon,
- Vorkommen des 3'-Heptamers innerhalb eines Bereiches von 10 kbp.

5 [0022] Die letzten drei Parameter können nur die Werte 0 oder 1 annehmen. Alle übrigen Parameter werden kontinuierlich bewertet. Primer, für die wenigstens ein Parameter mit 0 bewertet wurde, werden von den weiteren Berechnungen ausgeschlossen.

[0023] Für die Bewertung der Parameterwerte zwischen den Punkten A und B sowie C und D werden zwei Funktionen ausgewählt (siehe Fig. 1):

10 – für den Bereich [A, B]:
Fkt. 1:

$$15 y = \frac{1}{B-A} * x - \frac{A}{B-A}$$

Fkt. 2:

$$20 y = 0,5 * \cos \left[\pi \left(\frac{1}{B-A} * x - \frac{A}{B-A} + 1 \right) \right] + 0,5$$

– für den Bereich [C, D]:

Fkt. 3:

$$25 y = - \frac{1}{D-C} * x + \frac{D}{D-C}$$

30 Patentansprüche

1. Verfahren zur Ermittlung von Primer-Sequenzen und Primer-Paaren zur Erzeugung einer Vielzahl benachbarter oder überlappender PCR-Fragmente für eine gegebene DNA-Sequenz, mit den Schritten:

- Erstellung einer Liste mit einer Vielzahl von Teilabschnitten der gegebenen Sequenz, die aufgrund ihrer Länge als Primer ausgewählt werden können,
- Ermittlung der Werte von charakteristischen Parametern für jeden Teilabschnitt, wobei mindestens ein Parameter kontinuierlich bewertet wird, indem für den Parameter charakteristische Parameterwertebereiche erfasst werden, für die der jeweilige Parameter entweder als optimal oder als ungeeignet zur Erzeugung von PCR-Fragmenten gilt, und diese Parameterwertebereiche durch Übergangsbereiche getrennt werden, die durch Übergangsfunktionen repräsentiert werden, die stetig sind und einen streng monotonen Verlauf besitzen, wobei dem jeweiligen Parameterwert in den Übergangsbereichen eine Bewertung entsprechend der Größe der zugehörigen Übergangsfunktion zugeordnet wird,
- Ausschluss von Teilabschnitten, die in Bezug auf mindestens einen Parameter als ungeeignet zur Erzeugung von PCR-Fragmenten bewertet werden,
- Gesamtbewertung der übrigen Teilabschnitte durch gewichtete Summenbildung der den Parameterwerten bei der kontinuierlichen Bewertung zugeordneten Größen,

Auswahl der Teilabschnitte als Primer, die nach der Gesamtbewertung vorbestimmten Auswahlkriterien entsprechen, und

- Speicherung, Anzeige und/oder Ausgabe der ausgewählten Teilabschnitte als Primer-Sequenzen und Anordnung zu Primer-Paaren.

2. Verfahren gemäß Anspruch 1, bei dem als Parameter die Position des Teilabschnittes auf der Sequenz, die Länge des Teilabschnittes, der GC-Gehalt, die Schmelztemperatur, die 3'-Oligomehrstabilität, die 3'-Oligomehrhäufigkeit und/oder die Markierung ungeeigneter Sequenzbereiche für den Teilabschnitt ermittelt werden.

3. Verfahren gemäß Anspruch 2, bei dem Teilabschnitte mit Längen von weniger als 12 und mehr als 35 Nukleotiden als ungeeignet und mit Längen von 20 bis 24 Nukleotiden als optimal bewertet werden.

4. Verfahren gemäß Anspruch 2 oder 3, bei dem Teilabschnitte mit GC-Gehalten von weniger als 15% und mehr als 85% als ungeeignet und von 45% bis 55% als optimal bewertet werden.

5. Verfahren gemäß einem der Ansprüche 2 bis 4, bei dem Teilabschnitte mit Schmelztemperaturen von weniger als 30°C und mehr als 90°C als ungeeignet und im Temperaturbereich von 50°C bis 55°C als optimal bewertet werden.

6. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Übergangsfunktionen durch Winkelfunktionen, lineare Funktionen oder Überlagerungen aus diesen gebildet werden.

7. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Auswahl von Teilabschnitten als Primer in Bezug auf die Fragmentlänge derart erfolgt, dass bei einer vorbestimmten optimalen Länge von Teilabschnitten für einen zuerst ausgewählten Teilabschnitt auf dem einen Strang der Sequenz die Teilabschnitte auf dem anderen Strang als ungeeignet bewertet werden, wenn sie zu einer Fragmentlänge führen, die geringer als eine vorbestimmte Mindestlänge oder größer als eine vorbestimmte Maximallänge ist.

8. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Auswahl von Teilabschnitten als Primer derart erfolgt, dass für einen zuerst ausgewählten Teilabschnitt auf dem einen Strang der Sequenz die Teilabschnitte

DE 100 62 566 A 1

auf dem anderen Strang als ungeeignet bewertet werden, wenn beide Teilabschnitte eine Schmelztemperaturdifferenz besitzen, die größer als 10°C ist, wobei die Teilabschnitte als optimal bewertet werden, wenn die Schmelztemperaturdifferenz kleiner als 0,5°C ist.

9. Verfahren gemäß Anspruch 8, bei dem in einem Sequenzbereich, für den ein Teilabschnitt als Primer auf dem einen Strang der Sequenz ausgewählt werden soll, die 15 Primer mit den höchsten Gesamtbewertungen der einzelnen Primer vorausgewählt werden und zu diesen entsprechend der vorbestimmten optimalen Fragmentlänge jeweils die besten zugehörigen Teilabschnitte auf dem gegenüberliegenden Strang ausgewählt werden, wobei aus den so gebildeten Paaren von Teilabschnitten das Paar mit der höchsten Gesamtbewertung, die sich aus der Bewertung der einzelnen Teilabschnitte, der Schmelztemperaturdifferenz der beteiligten Primer und der Fragmentlänge ergibt, ausgewählt wird.

10. Verfahren gemäß Anspruch 9, bei dem als zusätzlicher Parameter in einem vorgegebenen Bereich der Sequenz die Lage des Primers relativ zu einem benachbarten Fragment zur Bildung einer vorbestimmten Mindestüberlappung erfasst wird.

11. Computerprogrammprodukt, das zur Ermittlung von Primer-Sequenzen und Primer-Paaren nach einem Verfahren gemäß einem der vorhergehenden Ansprüche eingerichtet ist.

12. Vorrichtung zur Ermittlung von Primer-Sequenzen und Primer-Paaren, die zur Durchführung eines Verfahrens mit den Schritten gemäß einem der vorhergehenden Ansprüche eingerichtet ist, mit:

- einer Einrichtung zur Erfassung einer Vielzahl von Teilabschnitten einer gegebenen DNA-Sequenz und von zugehörigen charakteristischen Parametern,
- einer Bewertungseinrichtung zur Bewertung der ermittelten Parameter,
- einer ersten Auswahleinrichtung, die dazu ausgelegt ist, als Primer ungeeignete Teilabschnitte der Sequenz zu verwerfen,
- einer zweiten Auswahleinrichtung, die dazu ausgelagert ist, die übrigen Teilabschnitte zu bewerten, und
- Einrichtungen zur Speicherung, Ausgabe und/oder Anzeige der ermittelten Primer-Sequenzen und Primer-Paaren.

Hierzu 1 Seite(n) Zeichnungen

5

10

15

20

25

30

35

40

45

50

55

60

65

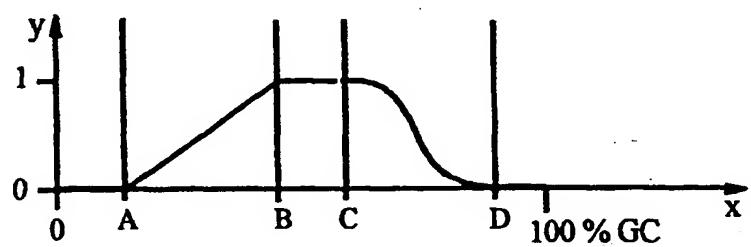


Fig. 1